Unstable regions in the scorecards' input space

Szűcs Imre

Szent István University, H-2103 Gödöllő, Páter Károly street 1,
icsusz@gmail.com

Abstract: Data mining algorithms become more and more popular to satisfy the Basle II requirements, like to predict the probability of default. Not all of these models can be understood easily from economical point of view, which involve the importance of stress tests. In this paper we try to map a retail credit scorecard's input space to find regions where predictions can lead to significant differing results. Different definitions for similarity and prediction difference are examined to reach an economically and statistically simultaneously interpretable abstraction.

Keywords: Basel II, retail credit risk, credit scorecard, data mining

1 Introduction

In the meaning of The New Basel Capital Accord (Basel II) [1] [2] Internal Rating Based (IRB) approach banks have to predict the probability of default (PD), loss given default (LGD) and exposure at default (EAD). For PD estimation the most widespread predictors are the logit regression based credit scorecards, but other data mining algorithms become more and more popular, like neural networks and support vectors machines [4] [5]. These methods suffer from the difficulties of business interpreting.

Besides developing the model itself, it is needed by Basel II to stress test the model from the economical volatility and stability point of view. To stress test the model a very important question has to be answered: for very similar applications how much the PDs could differ from each other? There is no exact definition on similarity and difference but its clear there is no chance to avoid the handling of these abstractions.

To examine the unpredictable regions of scorecards a logistic regression model - with logit link function - based application scorecard was developed to predict the probability of default of the credit agreements in one year time after the disbursement date. Then it was examined that for different similarity definitions how the prediction changes.

2 Material and method

For developing the credit scorecard a database was used with 3767 approved credit application. The database contained the application information and a target variable with the meaning of the default event in 1 year time after the disbursement. The aim of the research was to map the scorecard's space which needs larger and larger database in case of any new variable (later discussed). To reach a database with size can be handled a simple scorecard was developed with only four continuous input variable. Table 1 shows the input variables and the target variable of the scorecard.

Variable	Label	Role	Dimension	Type	Min	Max
Age	Age at application	Input	Year	Continuous	18	65
Emp	Employment years	Input	Year	Continuous	0	6
Dependents	Number of dependents	Input	Number	Continuous	0	6
Repayment	Monthly repayment amount	Input	HUF	Continuous	0	300 000
GB	Good / Bad flag	Target	Number	Binary	0	1

Table 1
Scorecard variables

The target variable means whether the application reached the default category in 1 year time after disbursement. In case an approved application got into default, the value of the target variable is 1, else 0.

The bad ratio in the database was: 23.47%.

The database was partitioned into a training (70%) and a Test (30%) sample randomly. The bad ratio in the training sample was 23.45% while in the test sample 23.50.

The continuous input variables were categorized based on the weight of evidence (WOE): The WOE shows the relative risk of the attribute.

$$WOE_{attribute} = \ln \frac{p_{attribute}^{nonevent}}{p_{attribute}^{event}} \tag{1}$$

where $p_{attribute}^{event} = \frac{n_{attribute}^{event}}{N^{event}}$ and $p_{attribute}^{nonevent} = \frac{n_{attribute}^{nonevent}}{N^{nonevent}}$,

and n_i^{event} and $n_i^{nonevent}$ means the BAD and GOOD applications number.

The variable selection was done by using the information value of the variable which calculated from the WOE of the variables' categories:

$$InformationValue = \sum_{attributes} \left[\left(p_{attribute}^{nonevent} - p_{attribute}^{event} \right) * WOE_{attribute} \right]. \quad (2)$$

After the 4 input variables was selected the categories' values was substituted by the weight of evidence of the given variable category.

Thus gave the base of the logistic regression with logit link function. The final scorecard had a 0.3728 KS (Kolmogorov Smirnov statistic) and 0 0.7438 AUR (Area under Receiving Operating Characteristic) value. [6] [7]

To map the scorecard behavior in different circumstances a larger and more detailed input space has to be created. Due to the limitations of computational capacity the space size was reduced by using different economically acceptable step sizes for the input variables. The age was stepped from 18 to 65 by 1, the employment from 0 to 6 by 0.1, the dependents from 0 to 6 by 1 and the repayment from 0 to 300 000 by 1000. In this 4 dimension input space for each points 10 other points was created inside a sphere with ε radius. The ε was defined differently for the different dimensions: ε_age=3, ε_employment=0.5, ε_dependents=1 and ε_repayment=5000 which has an economic mean that inside these values the applications can be seemed to be similar. Then for all points of the both the input space and the 10 close points the probability of default was calculated by the scorecard.

In this study the scorecard is reckoned as unstable at an input point if the prediction from the given input point and the prediction from the close points of the given point lead to sufficiently different results. This definition depends on the definition of „close" and the definition of „sufficiently differ". To measure the definition dependency, different values for them was examined. *Sufficiently* was analyzed as the difference between the predicted default probability of a given point and the predicted default probability of a close point. Then the prediction from *close* points was studied as how many from the 10 close points lead to different prediction.

3 RESULTS

The results depend consumedly on the applied definitions. Table 2 shows that while raising the problematic prediction distance and the percent of predictions have to be in the problematic the unstable region size decreasing. If we say that a prediction difference can be accepted only when the difference between the probability of defaults less then or equal to 0.05, and the model is unstable at an input point in case minimum 5 from the 10 close points are differ sufficiently from the centre point prediction, then table 2 shows that model is unstable in case of 0.76% of the input space. Thus means that in these regions the predicted probability of default could be change about 5% if the applicant is substituted with

another one, which is economically equal to the first one. To predict different future from economically equal points lead to inconsistent future. [8] Acceptable size of unstable region must be a business decision.

		Minimum number of outlying predictions									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Maximum distance between predictions	**0.01**	72,29%	63,72%	52,55%	40,30%	28,38%	17,87%	9,65%	4,23%	1,35%	0,24%
	0.02	49,05%	41,13%	31,58%	22,19%	14,16%	7,98%	3,76%	1,40%	0,38%	0,06%
	0.03	24,60%	16,84%	11,09%	6,99%	4,17%	2,29%	1,10%	0,43%	0,12%	0,02%
	0.04	12,83%	6,73%	3,78%	2,32%	1,42%	0,81%	0,40%	0,16%	0,05%	0,01%
	0.05	6,42%	3,55%	2,11%	1,29%	0,76%	0,40%	0,18%	0,07%	0,02%	0,00%
	0.06	3,90%	2,37%	1,46%	0,88%	0,48%	0,23%	0,09%	0,03%	0,01%	0,00%
	0.07	2,56%	1,57%	0,96%	0,56%	0,29%	0,13%	0,05%	0,01%	0,00%	0,00%
	0.08	1,70%	0,92%	0,50%	0,27%	0,13%	0,06%	0,02%	0,01%	0,00%	0,00%
	0.09	1,15%	0,56%	0,27%	0,13%	0,06%	0,02%	0,01%	0,00%	0,00%	0,00%
	0.10	0,79%	0,36%	0,17%	0,08%	0,03%	0,01%	0,01%	0,00%	0,00%	0,00%
	0.11	0,54%	0,21%	0,08%	0,03%	0,01%	0,00%	0,00%	0,00%	0,00%	0,00%
	0.12	0,40%	0,14%	0,04%	0,01%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	0.13	0,29%	0,10%	0,03%	0,01%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	0.14	0,20%	0,06%	0,02%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	0.15	0,11%	0,03%	0,01%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	0.16	0,07%	0,02%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	0.17	0,03%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	0.18	0,02%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	0.19	0,01%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	0.20	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%

Table 2

Instability in different cases

The unstable regions have to be analyzed deeply. Two type of answer could come up:

1. The model works properly, but there is some economically understandable reason that result the mentioned difference in a given region. In this case the *similarity* and / or the *difference* definition may need some modification.
2. This is a model specific error. In this case the aspects of the error have to known and handled. For example when an application form is filled out and the scoring solution recognizes that the input values are contained by the model unstable region then other judgment process is needed.

Summary

Measuring the PD model stability is an important part of the Basel II validation process. This is a must to examine how the model handles the economical volatility and the input variables variability. Both the capital requirement and the provision include PD as input parameter, which result the importance of predicting PD precisely. A 5% deviation in PD can lead to significant loss.

In this paper it was found that from behave of a given credit scorecard unstable regions of input space can be mapped. The importance of definitions was also pointed out. These regions can be used to attract attention on the weakness of the model, or to the specialties of the examined problem. In case of weakness further development of the model is needed or the idiosyncrasy of the prediction has to be handled.

Acknowledgement

I would like to express my gratitude to MKB Bank who gave me the possibility to complete this work by using their data. I have furthermore to thank SAS Institute Hungary who let me use SAS software to develop credit scorecard and to analyze it's behave. I am bound to my supervisor at Szent István University, Dr. Pitlik László for his valuable remarks on my PhD thesis.

Bibliography

[1] Basel Committee on Banking Supervision: Basel II: International Convergence of Capital Measurement and Capital Standards: a Revised Framework, 2004

[2] Basel Committee on Banking Supervision: The internal rating-based approach, Consultative document, January, 2001

[3] Basel Committee on Banking Supervision: Studies on Validation of Internal Rating Systems, 2005

[4] Imre Szűcs: Inconsistent Predictions by cross sell supporting behavior scorecards, in Proceedings of the Symposium for young researchers, Feast

of Hungarian Science, Budapest Tech, Hungary, November 3, 2006, pp. 215-224

[5] Hardle, Moro and Schafer: Estimating probability of default with support vector machines, Economic Risk, SFB Discussion Paper 2007-035, 2007

[6] Engelmann, Hayden and Tasche: Testing rating accuracy, www.risk.net, 2003

[7] Jorge Sobehart, Sean Keenan and Roger Stein: Validation methodologies for default risk models, Credit, May, 2000, pp. 51-56

[8] Pitlik, Pető, Pásztor, Popovics, Bunkóczi, Szűcs: Consistency controlled future generating models, EFITA, Vila Real, Portugal, 2005